Deutsches Institut für
Wirtschaftsforschung

DIW BERLIN

www.diw.de

The German
Socio-Economic
Panel Study

SOEPpapers
on Multidisciplinary Panel Data Research

372

Raphael Studer • Rainer Winkelmann

Specification and Estimation of Rating Scale Models - with an Application to the Determinants of Life Satisfaction

Berlin, April 2011

SOEPpapers on Multidisciplinary Panel Data Research
at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

The SOEPpapers are available at
http://www.diw.de/soeppapers

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: Uta Rahmann | soeppapers@diw.de

Specification and Estimation of Rating Scale Models – with an Application to the Determinants of Life Satisfaction*

Raphael Studer [a] and Rainer Winkelmann [a,b]

[a]*Department of Economics, University of Zurich*

[b]*CESifo, Munich, and IZA, Bonn*

March 2011

Abstract: Rating variables indicate the extent to which a quality is present, or absent, in a unit of observation. In this paper, we discuss a class of non-linear regression models for rating dependent variables and their estimation by parametric and semiparametric methods. An application to life satisfaction illustrates the main differences between the Rating Scale Model and ordinary least squares.

Keywords: rating variables, non-linear least squares, quasi-maximum likelihood, semiparametric least squares, subjective well-being

JEL Codes: C21, I00.

*Address for correspondence: University of Zurich, Department of Economics, Zürichbergstr. 14, CH-8032 Zürich, Switzerland, ☎ +41 44 634 22 97 and +41 44 634 22 92, raphael.studer@econ.uzh.ch and rainer.winkelmann@econ.uzh.ch. We thank Alois Stutzer for supporting our efforts to replicate results from Stutzer and Frey (2008).

1 Introduction

Empirical research using rating data has burgeoned in recent years. A rating variable represents the extent to which a quality (e.g., health, risk aversion, approval with a policy or party) is present, or absent, in a study unit. The rating is often, but not necessarily, coded on an integer-valued scale. The smallest value (commonly a zero) represents the complete absence of the quality, whereas the largest value represents its complete presence.

So far, regression analyses for such rating dependent variables have followed one of two approaches: Either, the rating is treated as an ordinal variable, indicating the use of ordered probit or ordered logit models. Or else, the rating is treated as cardinal and simple linear regression models are employed. The decision of the approach to follow rests in part on the number of categories. In fact, as pointed out by van Praag and Ferrer-i-Carbonell (2004), the distinction between the two blurs, and the distribution of the latent response index that underlies the ordered approach is fully identified, as the number of categories goes to infinity.

In this paper, we advocate an alternative approach for estimating the effects of explanatory variables on a rating, based on a class of non-linear single index regression models. As in linear regression, we focus on the conditional expectation as key object of interest. However, in order to maintain model consistency, we require that the conditional expectation respects the upper and lower bounds implied by the rating scale. As a consequence, predictions outside the range of the dependent variable are impossible and marginal effects are not constant. The model is easy to implement. It works for any number of categories, and extensions to panel data and instrumental variable estimation are feasible.

While the arguments developed in this paper apply to any regression with a rating dependent variable, we concentrate on a specific application, namely that of the economic determinants of self-rated well-being. Many household (panel) surveys include a single-item 7-point or 11-point question on general life satisfaction, as well as on satisfaction with various life domains (health, family, work etc.). In the previous literature either the linear regression model or ordered latent models have been used (or sometimes also both, see e.g., Clark and Oswald, 1996; Ferrer-i-Carbonell and Frijters, 2002; Frey and Stutzer, 2005).

The next section presents some further informal discussion of the pros and cons of various approaches to the regression analysis of rating scale variables. Section 3 provides a formal exposition of rating scale models. The methodology is illustrated in an application to the effect of time spent commuting to work on life satisfaction in Section 4. Section 5 concludes.

2 Motivation

Textbook treatments of rating variables routinely recommend the ordered probit and the ordered logit models (e.g., Cameron and Trivedi, 2005). These can be derived from a latent linear model with standard normally or logistically distributed errors, respectively, where a partition of the real line is used to generate the observed discrete distribution of ordered outcomes. In such models, the focus is on the probability distribution and its changes rather than on conditional expectations. The main advantage of ordered latent models is the implied conformity to the scaling of the rating dependent variable. In terms of the underlying latent linear variable, these models do not impose an equidistance between answer categories of the discrete scale.

However, although the name "ordered response model" suggests otherwise, the estimation method has a cardinal foundation as well (van Praag and Ferrer-i-Carbonell, 2004). For example, it is perfectly reasonable to make statements such as "the shift required to move a response from rating j to $j+1$ is twice as large as that required to move a response from $j+1$ to $j+2$". This raises the question, why a model with an implicit cardinalization should be preferred over a model which makes the cardinalization explicit.

In practice, these textbook models are therefore often abandoned in favor of the simpler linear regression model. Indeed, researchers on life satisfaction seem to have little discomfort in giving up the ordinal interpretation of the rating dependent variable and reporting mean satisfaction levels (for instance by country, or group; see e.g., Stone et al., 2010 and Sacks et al., 2010). If one follows these practitioners and accords plausibility to reported (conditional) mean rating values, the only factors speaking against the use of the linear regression model are that it imposes constant marginal effects and can predict rating scores outside the range of the rating scale.

The obvious remedy is to use a non-linear regression model that respects the boundaries of the rating dependent variable. If the attention is restricted to the class of single index models, the problem then becomes one of modeling the conditional expectation function (CEF) $\mathrm{E}(y|x) = G(x'\beta)$, where G is a twice differentiable monotonic function such that $y^{min} \leq G(x'\beta) \leq y^{max}$ for all values of x and β. If $y \in \{0, 1\}$ (the rating takes only two values), this model has the form of standard binary response models. This similarity is deceiving, though, because it is only in the binary response model that probability function and conditional expectation function coincide. For more than two-valued rating scales, the non-linear CEF model and the ordered response model constitute two truly different approaches.

We introduce such a rating scale model (RSM) and discuss the different assumptions regarding the G function in order to estimate the RSM. If a given parametric form is selected, estimation can proceed by non-linear least squares or quasi-maximum likelihood (see Papke and Wooldridge, 1996, for a closely related approach to fractional data). On the other hand, semiparametric least squares, introduced by Ichimura (1993), can be used in order to estimate the RSM without making functional form assumptions.

3 Econometric Rating Scale Model

3.1 Specification

A rating variable y has domain $y \in [0, y^{max}]$; where we have normalized the lower bound $y^{min} = 0$ for convenience. Thus the value "0" represents the complete absence of the quality, whereas y^{max} represents its complete presence. Suppose that there are N observations, and that y_i, $i = 1, \ldots, N$, is the rating for observation unit i.

The RSM is defined by a non-linear CEF

$$\mathrm{E}(y_i|x_i) = G(x_i'\beta) \tag{1}$$

such that $0 \leq G(.) \leq y^{max}$. The vector x_i is of dimension $(k \times 1)$ and β is a conformable parameter vector. The twice differentiable monotonic function $G(.)$ specifies the non-linear relationship between the additive linear index $x_i'\beta$ and the rating variable y_i. In a parametric RSM, a specific functional form is assumed for G. Two such possible specifications are a logit type model

$$G(x_i'\beta) = y^{max}\frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} \tag{2}$$

and a probit type model

$$G(x_i'\beta) = y^{max}\Phi(x_i'\beta) \tag{3}$$

where $\Phi(.)$ is the cumulative density function of the standard normal distribution. These models imply that the transformed rating scale $z_i = y_i/y^{max}$ has a standard logit- or probit CEF, respectively.

Specifications (2) and (3) guarantee that the CEF, as well as its prediction, always falls within the boundaries of the dependent variable. They also imply non-constant marginal effects. In the logit RSM

$$\frac{\partial E(y_i|x_i)}{\partial x_{il}} = y^{max}\frac{\exp(x_i'\beta)}{(1 + \exp(x_i'\beta))^2}\beta_l \tag{4}$$

In the probit RSM

$$\frac{\partial E(y_i|x_i)}{\partial x_{il}} = y^{max}\phi(x_i'\beta)\beta_l \tag{5}$$

where $\phi(.)$ denotes the density function of the standard normal distribution.

In such parametric frameworks, the model parameters can be estimated by non-linear least squares or by quasi-maximum likelihood, as explained in the next section. Alternatively, one can refrain from specifying the functional form of $G(.)$ and rather estimate it from data, together with the parameters β. This is a standard semiparametric estimation problem, and one possible estimator is due to Ichimura (1993).

3.2 Estimation

Re-writing model (1) as

$$y_i = G(x_i'\beta) + \epsilon_i$$

where ϵ_i is the CEF error and $\mathrm{E}(\epsilon_i|x) = 0$ by construction, it is easy to see that the model cannot be linearized. Alternativley, one could start from the model

$$y_i = G(x_i'\beta + \eta_i)\,, \;\; \mathrm{E}(\eta_i|x) = 0$$

in which case $G^{-1}(y_i) = x_i'\beta + \eta_i$. This approach has been proposed, in the context of a dependent variable bounded between 0 and 1, by Aitchison (1986). For a logit RSM, we can write

$$\log\left(\frac{y_i/y^{max}}{1 - y_i/y^{max}}\right) = x_i'\beta + \eta_i \tag{6}$$

There are two problems with this approach, however. First, y_i cannot take the extreme values of 0 or y^{max}. Second, it is impossible to recover the grandeurs of interest, especially the conditional expectation of the dependent variable y_i, since

$$\mathrm{E}(y_i|x_i) = \mathrm{E}\left(y^{max}\frac{\exp(x_i'\beta)\cdot\exp(\eta_i)}{1+\exp(x_i'\beta)\cdot\exp(\eta_i)}\middle|\, x_i\right) \neq y^{max}\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$$

Thus, model (6) is substantially different from (2) and as a consequence, it is hard to interpret β, the estimand in the linearized regression model (6), other than saying that β measures the effect of x on the logratios. To estimate the CEF parameters of the orginial non-linear RSM, a truly non-linear estimator is required, and we discuss non-linear least squares and quasi-maximum likelihood estimation in turn.

3.2.1 Non-linear least squares

Non-linear least squares minimizes the sum of squared residuals of model (1). This is equivalent to choosing $\hat{\beta}$, which solves the following first order condition:

$$s(\beta; y, x) = \sum_{i=1}^{N}(y_i - G(x_i'\beta))g(x_i'\beta)x_i = 0 \text{ where } g(x_i'\beta) = \frac{\partial G(x_i'\beta)}{\partial x_i'\beta} \tag{7}$$

As a member of the family of extremum estimators, the NLS estimator is consistent, if the sample is independent and identically distributed and if $G(.)$ fulfills some regularity conditions (e.g., Hayashi, 2000).

Asymptotic theory enables the computation of standard errors. Default options in statistical software packages assume a spherical error variance. However, due to the boundedness of a rating variable the variance is heteroscedastic. Intuitively, the closer the rating score moves to the boundaries the less dispersion is possible. The error term $\varepsilon_i = y_i - E(y_i|x_i'\beta)$ inherits the heteroscedasticity of the rating variable. Therefore, a heteroscedastic consistent variance-covariance estimator for $\hat{\beta}$, as proposed by Huber (1967) and White (1980) is employed:

$$\text{AVar}(\hat{\beta}) = n^{-1}I^{-1}(\beta)J(\beta)I^{-1}(\beta)$$

where

$$I(\beta) = \text{E}[H(\beta; y, x)] = \text{E}\left[-g(x_i'\beta)^2 x_i x_i'\right]$$

and

$$J(\beta) = \text{Var}(s(\beta; y, x)) = E\left[(y_i - G(x_i'\beta))^2 g(x_i'\beta)^2 x_i x_i'\right]$$

Replacing the population moments reported above by their sample analogs leads to a consistent estimator of the heteroscedastic consistent variance-covariance matrix of $\hat{\beta}$.

3.2.2 Quasi-Maximum Likelihood Estimation

The parameters of the RSM (1) can be estimated consistently by embedding it in any member distribution of the linear exponential family and using maximum likelihood. Available distributions include, among others, the normal distribution, the Poisson distribution and the Bernoulli distribution (Gourieroux et al., 1984). The only requirement for consistency is that the CEF of the RSM is correctly specified. This approach is referred to as quasi-maximum likelihood estimation (QML).

For example, if the normal distribution is used, QML is equivalent to non-linear least squares. If the Bernoulli distribution $B(1,p)$ is used as a basis for estimation, one needs to observe that $0 \leq p \leq 1$, whereas the CEF of the RSM is bounded from above at y^{max}. This problem can be solved by dividing both sides of equation (1) by y^{max}. The Bernoulli QML estimator is obtained by setting $p_i = G(x_i'\beta)/y^{max}$, and the first order conditions are:

$$\sum_{i=1}^{N} \frac{y_i - G(x_i'\beta)}{y^{max}} \frac{g(x_i'\beta)}{(1 - G(x_i'\beta)/y^{max})G(x_i'\beta)} x_i = 0 \text{ where } g(x_i'\beta) = \frac{\partial G(x_i'\beta)}{\partial x_i'\beta} \tag{8}$$

The QML framework does not impose any restrictions on the second or any higher moment of the dependent variable. In fact, the second moment is misspecified in the Bernoulli QML framework. Hence, the maximum likelihood variance estimation, which equals the inverse of the Hessian's expectation, has to be replaced by the robust sandwich variance estimator (Gourieroux et al. 1984).

3.2.3 Comparison and Implementation

For a correctly specified CEF, both NLS and Bernoulli QML are consistent estimators. In small samples they may differ, since they use different weights w_i for the sample analog of the set of orthogonality conditions:

$$\sum_{i=1}^{N}(y_i - G(x_i'\beta))x_i w_i = 0 \tag{9}$$

On one hand, NLS weighs the orthogonality conditions with the standard normal or the logistic probability density functions, respectively. On the other hand, the Bernoulli QML estimator weighs observations with the probability density divided by the variance of a Bernoulli distributed variable. For the logistic model, these terms cancel and all elements of the score vector are weighted equally. The optimal weighting scheme depends on the true data generating process and its higher order moments. Since no such assumptions were made in our rating scale model, estimation with equal weights appears like a good starting point.

Both estimation methods are easy to implement in standard statistical software packages. In Stata (StataCorp., 2003), for example, the relevant model environment is given by the generalized linear model (glm) command. It allows to define distribution as well as link function. Choosing the normal distribution in conjunction with the logit link gives, for example, the non-linear least squares estimators of the logit-type RSM. Choosing the Bernoulli distribution instead results in the corresponding QML estimator. In either case, all ratings have first to be divided by the upper bound y^{max}, and robust standard errors need to be computed.

3.2.4 Semiparametric Least Squares

NLS and Bernoulli QML provide consistent parameter estimates for model (1) if the conditional expectation is correctly specified. Alternatively, one estimate the $G(.)$-function jointly with the regression parameters β. This approach remains consistent for β as long as the single index structure holds, regardless of the true $G(.)$. Different semiparametric estimators can be used. This paper employs one that does not rely on higher order moment conditions and that is the most simple to implement, namely semiparametric least squares

(SLS) proposed by Ichimura (1993). SLS minimizes the sum of squared residuals of model (1).

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \hat{\mathrm{E}}(G(x_i'\beta)|x_i'\beta))^2 \tag{10}$$

Iterative methods with an initial guess on $\hat{\beta}$ have to be applied in order to estimate both β and $\mathrm{E}(G(x_i'\beta)|x_i'\beta)$. For the latter, the local constant estimator proposed by Nadaraya (1965) and Watson (1964) is used. The local constant estimator depends on a kernel function and a bandwidth. If the choice of the kernel does not matter much, the bandwidth selection is important. The most appropriate way to choose the optimal bandwidth in kernel regression is to apply cross validation (see e.g., Cameron and Trivedi, 2005).

Assuming an independent and identically distributed sample, a bandwidth sequence which converges towards 0 as N increases, as well the validity of some technical conditions on parameter space and kernel, it is possible to show that the SLS estimator the is consistent and asymptotically normal (Ichimura, 1993). Parameters are identified only up to location and scale. In other words, any additive and multiplicative shifts in the regressors are incorporated by $G(.)$. Therefore, x_i does not include a constant term, and all remaining parameters are normalized with respect to the parameter of a continuous regressor. Marginal effects can be recovered for all explanatory variables, and standard errors can be bootstrapped.

The semiparametric RSM can be implemented conveniently using the non-parametric package in R (Hayfield and Racine, 2008). The program routine chooses the optimal bandwidth using cross validation and proposes as outputs estimates of the parameter vector, marginal effects and bootstrapped standard errors for those estimates.

4 Empirical Application to Life Satisfaction Data

Stutzer and Frey (2008), in their paper "Stress that doesn't pay: The commuting paradox", analyzed the effect of commuting time on satisfaction using linear regression models. We replicate one of their analyses and re-estimate it using the rating scale model proposed in the previous sections.

Here, the rating dependent variable *"overall live satisfaction"*, measured on a discrete scale ranging from 0 to 10, was modeled. The explanatory variable of interest, "one way commuting time to work", was measured in minutes. Data from eight waves of the German Socio Economic Panel (Wagner et al., 2007) (1985, 1990, 1991, 1992, 1993, 1995, 1998, 2003) were used. The sample excluded people with irregular commuting patterns. Commuting times for people working from home were set to zero. The authors pooled all eight waves and estimated the model by OLS. These estimates can be found in column 3 of table 1 in Stutzer and Frey (2008). Even though the number of observations used in this replication differs from that used by Stutzer and Frey by 707 observations, summary statistics, which are reported in Table 3 in the appendix, and linear regression estimates are virtually the same.

4.1 Parametric estimation results

Table 4 in the appendix reports OLS and parametric RSM estimates of the parameter vector β from model (1) specifying the set of explanatory variables as proposed by Stutzer and Frey (2008). These include gender, age, age^2, 6 categories of years of education, 2 variables for the relationship to the household head, 9 variables for marital status, 4 variables for number of children in the household, the square root of the number of household members, East German, foreigners with EU nationality, foreigners without EU nationality, and self-employment, in addition to the key variable of interest, commuting time.

The estimated average marginal effects of a one-minute increase in commuting time on life satisfaction scores are shown in Table 1 below (with standard errors of in parentheses). Column 1 replicates the ordinary least squares estimates found by Stutzer and Frey (2008). Column 2 to 5 report average marginal effects of the parametric RSM. In columns 2 and 3 the Bernoulli QML estimates are shown. NLS estimates are given in columns 4 and 5, respectively.

OLS predicts the highest average reduction in satisfaction scores. A person commuting 60 minutes one-way is expected to have a 0.275 point lower satisfaction score than a comparable person, who does not commute (Stutzer and Frey 2008). But the average effects from the other models are very similar. For example, the effect obtained by the logit-type Bernoulli QML estimation amounts to a 0.269 point decrease. It appears that for this application, the different weighting schemes employed by the NLS and Bernoulli QML estimators for the parametric RSM do not matter much.

Table 1: Average Marginal Effect of Commuting Time (in minutes) on Satisfaction

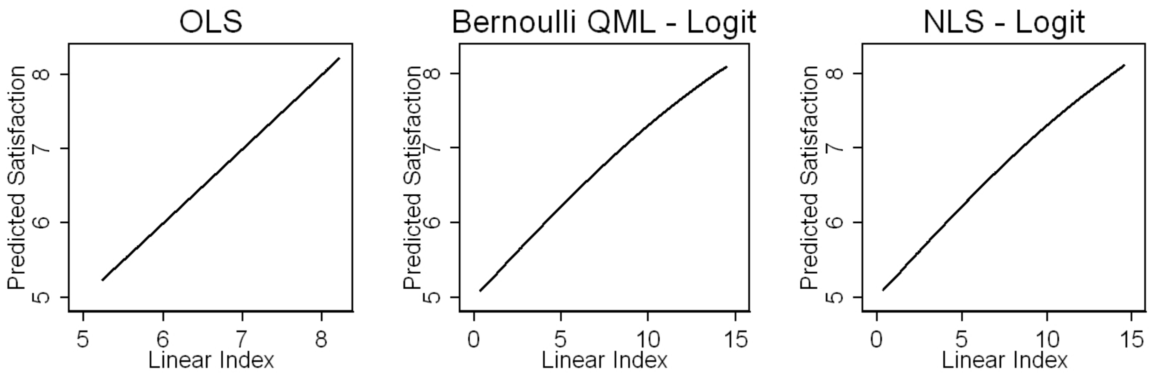
	(1) OLS	(2) QML-Logit	(3) QML-Probit	(4) NLS-Logit	(5) NLS-Probit
Commuting Time	-0.00459	-0.00449	-0.00453	-0.00451	-0.00453
	(0.00046)	(0.00047)	(0.00047)	(0.00047)	0.00048
Individual characteristics	Yes	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Robust Standard Errors	No	Yes	Yes	Yes	Yes
Observations	39747	39747	39747	39747	39747

· Robust standard errors in parentheses.
· Column (1) corresponds to column 3 of table 1 in Stutzer and Frey (2008).

In the light of the wide utilization and acceptance of OLS in the rating variable literature, it is appealing to find the parametric RSM estimate very similar average marginal effects. However, the non-linear specification of the conditional expectation of the rating dependent variables has two main implications distinguishing the RSM from OLS. These points and the resulting superiority of the RSM will be highlighted in graphical illustrations.

Figure 1 plots the estimated conditional expectation, i.e. predicted satisfaction scores for all sample members. The three graphs report the predictions obtained by OLS, Bernoulli QML and NLS (from left to right). The latter two models are based on the logit specification. For OLS, the mean predictions are simply equal to the linear index. In this application, OLS predictions are far away from the bounds set by the response scale, here 0 and 10, and extreme out-of-sample predicitions would be required to cause a violation

Figure 1: In-sample predictions of satisfaction scores



of the bounds to occur. But, this need not hold in general, and the approach is a-priori model inconsistent.

The Bernoulli QML and the NLS predictions are very similar. Both predict a locally concave relationship between linear indexes and life satisfaction scores. Hence, the sample predictions are centered around the upper flection of the logistic cumulative distribution function.

Figure 2: Marginal Effect of Commuting Time (in minutes) on Satisfaction

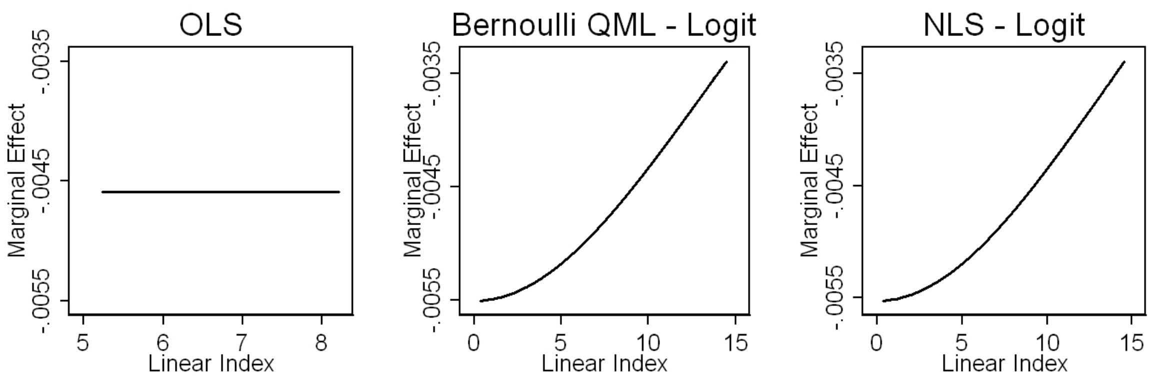


Figure 2 plots the estimated marginal effects for all sample members. For OLS the marginal effect is constant among all individuals. In the second and third graph of Figure 2 the individual specific marginal effects are shown for the logit type Bernoulli QML and the NLS models. The graphs suggest that commuting time affects people in the tails of the distribution of predicted satisfaction scores by over 20% more or less than estimated by OLS. Moreover, we find that with an increasing linear index commuting time affects individuals less. This is plausible. Very satisfied people, who feel themselves fully blessed with luck, weigh a one-minute increase in commuting time less than people, who perceive their life as unsatisfactory. The non-constant marginal effects are therefore a useful feature, although they are model driven, and not identified from the data. This is what the semi-parametric results presented in the next section do.

4.2 Semiparametric estimation results

We choose to implement the SLS estimator using a plug-in bandwidth for two reasons. First, the huge sample and the big number of parameters make cross validation computationally intensive. Second, several tries in random subsamples showed that cross validation chose the bandwidth too small, that resulted in an under-smoothed estimate of the conditional expectation of $G(.)$. This might be due to the lack of independence among observations, as the sample is pooled over time periods. Different essays identified a plug-in bandwidth of 10 to provide appropriate smoothing.

Table 2: Average Marginal Effect of Commuting Time (in minutes) on Satisfaction

	(1) OLS	(2) SLS
Commuting Time	-0.00459	-0.00479
Individual characteristics	Yes	Yes
Time fixed effects	Yes	Yes
Observations	39747	39747

· Marginal effects in column (2) are evaluated at the mean characteristics.
· Column (1) corresponds to column (3) of table 1 in Stutzer and Frey (2008).

Column 1 of Table 2 shows the same OLS marginal effect as presented in column 1 of Table 1. Column 2 of Table 2 reports the marginal effect of commuting time on life satisfaction estimated by SLS and evaluated at the mean characteristics. SLS and OLS estimates are very similar. A 30 minutes increase in one-way commuting time lowers life satisfaction approximately by 0.144 respectively 0.138 point ceteris paribus. Table 5 in the appendix shows that this finding holds for the marginal effects of all explanatory variables.

Figure 3: Predicted Satisfaction for Sample Members

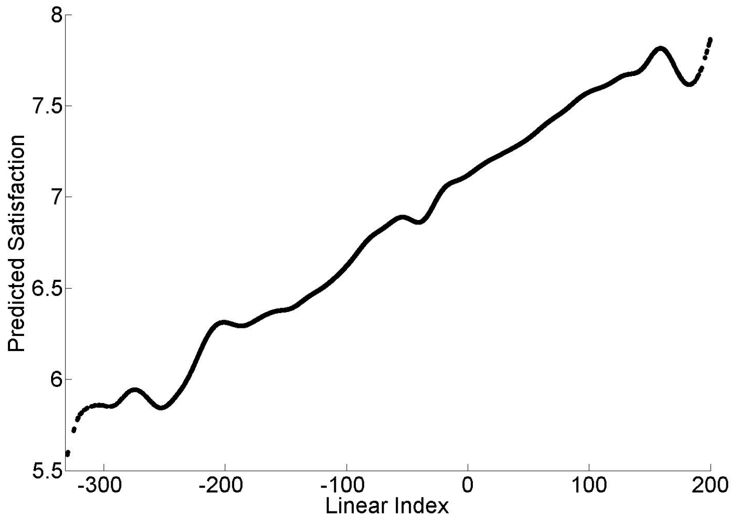


Figure 3 plots the SLS mean predictions. The range of predicted satisfaction scores is similar to the OLS predictions. The main difference of Figure 3 to Figure 1 is the widespread linear index, which is due to the small normalization parameter (the coefficient of commuting time). A last peculiarity deserves to be mentioned. For being in line with Figure 1, we changed the sign of the linear index in Figure 3. In fact, relative coefficients take the opposite sign of the OLS coefficients, as the normalization parameter is negative. Hence, the untransformed estimated CEF would actually be decreasing in the linear index.

Several concluding remarks apply. First, the SLS estimates of marginal effects are very close to OLS (and therefore to the average marginal effects of the parametric RSM). Second, SLS respects the boundaries of the rating dependent variable as the observed data is used to estimate the conditional expectation of $G(.)$. Finally, researchers should be aware that SLS does not allow for out-of-sample predictions.

5 Conclusion

This paper focuses on econometric models for rating data. Existing models, such as ordered latent models or the linear regression model, have a number of shortcomings. A new general framework for a cardinal rating scale model addresses these issues. Depending on the specific assumptions, model parameters can be estimated by non-linear least squares, by quasi-maximum likelihood or by semiparametric least squares.

Predicted means of these rating scale models automatically satisfy the logical constraints provided by the upper and lower bounds of the scale. They work equally well for discrete ratings, as for continuous ones. An example for a near continuous rating scales are the Standard & Poors ratings of investment grades, that distinguishes 25 values. Truly continuous ratings are also possible, by representing them as points on a line. For instance, degrees of approval or disapproval can be elicited by asking subjects to position a visual mark on a ruler. This method has been employed occasionally in psychometrics, and is likely to become more widespread in the future. In these cases, ordered latent models are clearly impractical, and the proposed RSM is a superior alternative to the linear regression model that ignores the intrinsic features of the underlying scale.

In an empirical application to discrete life satisfaction scores illustrated the implementation of these methods in a concrete empirical setting. It turned out that the average marginal effects of the nonlinear RSM were similar to ordinary least squares estimates. However, substantial differences in predicted individual specific marginal effects could be found for observations in the tails of the distribution of predicted satisfaction scores.

References

Aitchison, J., 1986, *The Statistical Analysis of Compositional Data*, Chapman and Hall, New York

Cameron, C.A. and P.K. Trivedi, 2005, *Microeconometrics, Methods and Applications*, Cambridge University Press, New York

Clark, A.E. and A.J. Oswald, 1996, "Satisfaction and Comparison Income', *Journal of Public Economics*, Vol. 61, Issue 3, 95-144

Easterlin, R. A., 1974, "Does Economic Growth Improve the Human Lot?", *David, P.A. and M. W. Reder, eds., Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, New York: Academic Press, Inc., 98-125

Ferrer-i-Carbonell, A. and P. Frijters, 2004, "How important is Methodology for the Estimates of the Determinants of Happiness?", *The Economic Journal*, Vol. 114, Issue 497, 641-659

Frey, B.S. and A. Stutzer, 2002, *Happiness and Economics*, Princeton University Press, Princeton

Frey, B.S. and A. Stutzer, 2005, "Happiness Research: State and Prospects", *Review of Social Economy*, Vol. 62, 207-228

Gourieroux C., A. Monfort and A. Trognon, 1984, "Pseudo Maximum Likelihood Methods: Theory", *Econometrica*, Vol. 52 No.3, 681-700

Hayashi, F. 2000, *Econometrics*, Princeton University Press, New Jersey

Hayfield, T. and J.S. Racine, 2008, " Nonparametric Econometrics: The np Package", Journal of Statistical Software 27(5). URL http://www.jstatsoft.org/v27/i05/

Huber, P. J., 1967, ”The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions”, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. I, pp. 221-233.

Ichimura, H., 1993, ”Semiparametric least squares (SLS) and weighted SLS estimation of singe-index models”, *Journal of Econometrics*, 58, 71-120

Kristoffersen, I., 2010, ”The Metrics of Subjective Wellbeing: Cardinality, Neutrality and Additivity”, *The Economic Record*, Vol. 86, No. 272, 98-123

Nadaraya, E.A., 1965, ”On nonparametric estimates of density functions and regression curves”, *Theory of Applied Probability*, 10, 186-190

Papke, L.E. and J.M. Wooldridge, 1996, ”Econometric methods for fractional response variables with an application to 401(k) plan participation rates”, *Journal of Applied Econometrics*, Vol. 11, Issue 6, 619-632

Racine, J.S, 2008, ”Nonparametric Econometrics: A Primer”, *Foundations and Trends in Econometrics*, 3(1), 1-88

Sacks, D.W., B. Stevenson and J. Wolfers, 2010, ”Subjective Well-being, Income, Economic Development and Growth”, *NBER Working Paper Series*, No. 16441

StataCorp, 2003, *Stata Statistical Software: Release 10*, College Station, Texas

Stone, A.A., J.E., Schwartz, J.E., Brodericka and A. Deaton, 2010, ”A Snapshot of the Age Distribution of Psychological Well-being in the United States”, PNAS Paper

Stutzer, A. and B.S. Frey, 2008, ”Stress that Doesn’t Pay: The Commuting Paradox”, *Scandinavian Journal of Economics*, Vol. 110, Issue 2, 339-366

The MathWorks Inc., 2008, *Matlab, The Language of Technical Computing: Version 7.6.0.324 (R2008a)* van Praag, B.M.S.and A. Ferrer-i-carbonell, 2004, *Happiness Quantified: A Satisfaction Calculus Approach*, Oxford University Press , New York

Wagner, G.G., J.R. Frick and J. Schupp, 2007, "The German Socio-Economic Panel Study (SOEP) - Scope, Evolution and Enhancements", *Schmollers Jahrbuch*, 127, 1, 139-169

Watson, G.S., 1964, "Smooth regression analysis", *Sankhya*, 26(15), 359-372

White, H., 1980, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica*, 48, 817-838

Appendix

Table 3: Replication of Summary Statistics

	Mean	Std.Dev.	Min	Max
Age	38.84	11.60	14	86
Years of Education	11.41	3.08	7	18
Children in hh	0.75	0.98	0	9
Persons in hh	3.12	1.35	1	14
Female	0.44	0.49	0	1
Child of hh-head	0.13	0.33	0	1
No hh-head	0.01	0.1	0	1
Single-wp	0.25	0.43	0	1
Married	0.65	0.48	0	1
Seperated-wp	0.02	0.13	0	1
Seperated-np	0.002	0.04	0	1
Divorced-wp	0.06	0.24	0	1
Divorced-np	0.004	0.07	0	1
Widowed-wp	0.01	0.12	0	1
Widowed-np	0.001	0.04	0	1
Spouse abroad	0.002	0.04	0	1
Selfemployed	0.15	0.36	0	1
East-German	0.2	0.40	0	1
EU-citizen	0.07	0.26	0	1
Foreigner Non-EU	0.1	0.3	0	1

· N=39747
· Abbreviations: hh: household, np: no partner, wp: with partner
· This table replicates the summary statistics provided in the appendix of Stutzer and Frey (2008).

Table 4: Raw Regression Output - Parametric RSMs

	OLS	QML-Logit	QML-Probit	NLS-Logit	NLS-Probit
Commuting Time$\cdot 10^{-2}$	-0.459	-0.220	-0.133	-0.221	-0.133
	(0.046)	(0.023)	(0.014)	(0.023)	(0.014)
Age $\cdot 10^{-2}$	-4.895	-2.441	-1.454	-2.503	-1.491
	(0.576)	(0.300)	(0.179)	(0.302)	(0.180)
Age^2 $\cdot 10^{-2}$	0.051	0.025	0.015	0.026	0.016
	(0.007)	(0.004)	(0.002)	(0.004)	(0.002)
Female	-0.018	-0.009	-0.005	-0.008	-0.005
	(0.017)	(0.008)	(0.005)	(0.008)	(0.005)
Education = 7y.	-0.042	-0.021	-0.013	-0.018	-0.011
	(0.046)	(0.025)	(0.015)	(0.025)	(0.015)
Education = 10y.	0.155	0.077	0.046	0.077	0.046
	(0.026)	(0.014)	(0.008)	(0.014)	(0.008)
Education = 12y.	0.194	0.096	0.058	0.096	0.057
	(0.033)	(0.016)	(0.010)	(0.016)	(0.010)
Education = 14y.	0.247	0.122	0.073	0.124	0.0743
	(0.037)	(0.018)	(0.011)	(0.018)	(0.011)
Education = 18y.	0.394	0.195	0.117	0.195	0.117
	(0.039)	(0.019)	(0.011)	(0.019)	(0.011)
Child of hh-head	0.086	0.043	0.027	0.037	0.023
	(0.043)	(0.021)	(0.013)	(0.021)	(0.013)
No hh-head	-0.168	-0.081	-0.050	-0.077	-0.047
	(0.084)	(0.043)	(0.026)	(0.043)	(0.026)
Single-wp	0.926	0.415	0.254	0.415	0.254
	(0.207)	(0.109)	(0.068)	(0.110)	(0.068)
Married	1.140	0.518	0.316	0.514	0.314
	(0.206)	(0.109)	(0.068)	(0.110)	(0.068)
Separated-wp	0.504	0.224	0.137	0.224	0.137
	(0.216)	(0.114)	(0.071)	(0.115)	(0.071)
Separated-np	-0.508	-0.221	-0.137	-0.217	-0.134
	(0.220)	(0.128)	(0.079)	(0.129)	(0.080)
Divorced-wp	0.769	0.345	0.211	0.342	0.210
	(0.209)	(0.110)	(0.068)	(0.111)	(0.069)
Divorced-np	-0.002	-0.007	-0.004	-0.005	-0.002
	(0.130)	(0.070)	(0.042)	(0.070)	(0.042)
Widow-wp	0.809	0.364	0.222	0.361	0.221
	(0.217)	(0.114)	(0.071)	(0.115)	(0.071)
Widow-np	-0.453	-0.203	-0.124	-0.201	-0.123
	(0.238)	(0.151)	(0.093)	(0.150)	(0.093)
Child-hh=1	-0.064	-0.031	-0.019	-0.031	-0.019
	(0.025)	(0.012)	(0.007)	(0.012)	(0.007)
Child-hh=2	-0.077	-0.038	-0.023	-0.036	-0.022
	(0.033)	(0.016)	(0.010)	(0.016)	(0.010)
Child-hh>3	-0.222	-0.109	-0.065	-0.110	-0.066
	(0.051)	(0.025)	(0.015)	(0.025)	(0.015)
Squareroot Persons in hh	0.111	0.055	0.033	0.056	0.033
	(0.040)	(0.020)	(0.012)	(0.020)	(0.012)
Selfemployed	-0.090	-0.044	-0.027	-0.044	-0.026
	(0.023)	(0.011)	(0.007)	(0.011)	(0.007)
East-German	-0.713	-0.336	-0.204	-0.336	-0.203
	(0.022)	(0.010)	(0.006)	(0.010)	(0.006)
EU-citizen	0.126	0.065	0.039	0.063	0.038
	(0.035)	(0.019)	(0.011)	(0.019)	(0.011)
Foreigner Non-EU	-0.119	-0.059	-0.035	-0.059	-0.035
	(0.030)	(0.016)	(0.010)	(0.016)	(0.010)
First interview	0.254	0.131	0.078	0.131	0.078
	(0.037)	(0.019)	(0.011)	(0.019)	(0.011)
Year 90	0.084	0.043	0.026	0.045	0.027
	(0.028)	(0.015)	(0.009)	(0.015)	(0.009)
Year 92	-0.412	-0.181	-0.111	-0.179	-0.110
	(0.067)	(0.029)	(0.018)	(0.029)	(0.018)
Year 95	-0.060	-0.030	-0.018	-0.026	-0.016
	(0.025)	(0.012)	(0.007)	(0.013)	(0.007)
Year 98	-0.011	-0.006	-0.004	-0.004	-0.003
	(0.026)	(0.013)	(0.008)	(0.013)	(0.008)
Year 03	-0.072	-0.036	-0.022	-0.032	-0.020
	(0.023)	(0.012)	(0.007)	(0.012)	(0.007)
Constant	7.110	0.947	0.578	0.959	0.585
	(0.241)	(0.126)	(0.078)	(0.127)	(0.078)

· Standard errors reported in parentheses.
· Estimated coefficients correspond to the parameter vector β in model (1).
· First line of column (1) corresponds to column (3) of table 1 in Stutzer and Frey (2008).
· N=39747

Table 5: Marginal Effects - Semiparametric RSM

	OLS	SLS
Commuting Time $\cdot 10^{-2}$	-0.459	-0.479
Age $\cdot 10^{-2}$	-4.895	-5.083
Age^2 $\cdot 10^{-2}$	0.051	0.053
Female	-0.018	-0.019
Education = 7y.	-0.042	-0.044
Education = 10y.	0.155	0.162
Education = 12y.	0.194	0.204
Education = 14y.	0.247	0.261
Education = 18y.	0.394	0.409
Child of hh-head	0.086	0.089
No hh-head	-0.168	-0.177
Single-wp	0.926	0.979
Married	1.140	1.175
Separated-wp	0.504	0.526
Separated-np	-0.508	-0.536
Divorced-wp	0.769	0.785
Divorced-np	-0.002	-0.002
Widow-wp	0.809	0.857
Widow-np	-0.453	-0.474
Child-hh=1	-0.064	-0.066
Child-hh=2	-0.077	-0.081
Child-hh>3	-0.222	-0.235
Squareroot Persons in hh	0.111	0.113
Selfemployed	-0.090	-0.095
East-German	-0.713	-0.743
EU-citizen	0.126	0.133
Foreigner Non-EU	-0.119	-0.125
First interview	0.254	0.266
Year 90	0.084	0.083
Year 92	-0.412	-0.429
Year 95	-0.060	-0.063
Year 98	-0.011	-0.012
Year 03	-0.072	-0.075
Constant	7.110	

- Reported coefficients correspond to marginal effects. Marginal effects in Column 2 are evaluated at the mean characteristics.
- First line of column (1) corresponds to column (3) of table 1 in Stutzer and Frey (2008).
- The life satisfaction score is modeled as dependent variable.
- N=39747